Comparative Wavelet and MFCC Speech Emotion Recognition Experiments on the RAVDESS Dataset

[1]Aayush Bajaj, [2]Abhishek Jha, [3]Lakshay Vashisth, [4]Dr. K.C. Tripathi

Department of Information technology, Maharaja Agrasen Institute of Technology, New Delhi, India

[1]dev.aayushbajaj@gmail.com,[2] jhabhishek3797@gmail.com, [3]lakshaylv97@gmail.com, [4]kctripathi@mait.ac.in

Article Info
Page Number: 1288-1293
Publication Issue:
Vol. 71 No. 3 (2022)

Article History
Article Received: 12 January 2022
Revised: 25 Febuary 2022
Accepted: 20 April 2022
Publication: 09 June 2022

Abstract— Emotion Recognition (ER) from speech is one of the most interesting research domains for the scientific world. The challenge behind ER is essentially the method of speech-feature-extraction that can efficiently encapsulate speaker independent emotional information from speech signals. This paper compares the performance of Window-Fourier Transform Method, Mel-Frequency Cepstral Coefficients (MFCC's) and Continuous/Discrete Wavelet Transforms from the perspective of constant vs variant localization of time-frequency on The Rayerson audio-visual database of emotional speech and song. Wavelet transform has proven to be a promising non-linear tool for signal analysis that has been successfully applied in image recognition, compression and other tasks. MFCC's has been a standard in feature extraction for speech. The motive here is to compare both the methods using the Random Forest algorithm with similar hyperparameters.

Index Terms—Continuous wavelet transform, Discrete Wavelet transform, Emotion recognition, Mel-Frequency Cepstral Coefficient

I. INTRODUCTION

The main motivation behind this project is to study the performance of non-linear speech analysis methods against the linear methods in speech emotion recognition for non-stationary signals. We selected **wavelet as the non-linear tool and MFCC as linear tool. The difference between Fourier and wavelet transforms from the perspective of time– frequency analysis is the localization methods of the two transforms.** The Fourier transform offers constant and uniform time–frequency resolution whereas the wavelet transform enables better frequency resolution at low frequencies and better time localization of the transient phenomena in the time domain [2]. This resembles to the first stage of human auditory perception [3] and to basilar membrane excitation [4]. Some notable works in this area includes performance comparisons on Automatic Speech Recognition in [5].

II. ER SYSTEM AND FEATURE EXTRACTION METHODS

A. RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech

includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). Note, there are no song files for Actor_18. To reduce complexity we've downsampled the dataset from 48khz to 16khz and retained only male voices corresponding to the transcript "Dogs are sitting by the door".[6]

B. Continuous Wavelet Transform

Continuous Wavelet Transform of a signal f(t)is given by:

$$W_{\psi(a,b)}[f] = \int_{-\infty}^{+\infty} f(t).\overline{\psi_{a,b}}\left(\frac{t-a}{b}\right) dt \qquad (1)$$

Where aR is the scaling factor and bR+ is the translation factor and (t) is a continuous function in time and frequency domain called the Mother wavelet and (t) is the complex conjugate. For different values of a,bR a mother wavelet produces many daughter wavelets.

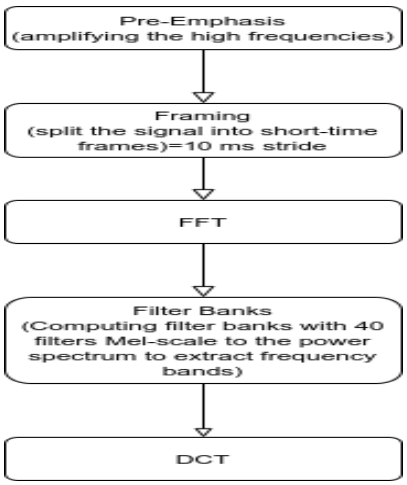
For the sake of this comparison we've chosen Morlet Wavelet due its close properties with human speech perception. The feature extraction with continuous morlet wavelet transform discussed in [7] is implemented for the sake of comparison in this paper.

C. Discrete Wavelet Transform

Discrete wavelet transform is implemented as a series of Band Pass filters (high and low). DWT of the signal discussed in [8] is implemented in this paper. Approximation coefficients corresponding to the output of low pass filter at every level and Detail coefficients corresponding to the high pass filter of the DWT are taken into account. The mother wavelet chosen for discrete transform is Daubechies [9] with 28 filter length.

D. MFCC

Mel-Frequency Cepstral coefficients were calculated according to the following procedure in [10]



E. Principal Component Analysis

The matrix size for CWT Wmxn (number of scales x length of speech signal), DWT W1xn(level of decomposition in filter banks x 1) and MFCC Wnxm (number of filters x time(depends on hop length)) are large. We needed some compact representation without losing any information regarding emotions.

Dimensional reduction methods, such as PCA method, are used in this paper to convert the original set of features to a different and more compact representation keeping as much information as possible and to try to increase the performance of the speech emotion recognition system. The process is discussed in detail in [10]

F. Gini Index

To calculate the quality of the split we've used the Gini impurity method.

$$G = \sum_{i=1}^{nclasses} p[i].(1 - p[i]) \quad (2)$$

p(i) is the probability of picking a datapoint with class (i)

produces many daughter wavelets.

G. Entropy

Calculate the entropy of distribution for given probability values of different extracted coefficients. Entropy is calculated from the following formula:

$$E(x) = -\sum_i x_i.\log(x_i) \quad (3)$$

H. Zero Crossing Rate

The Zero-Crossing Rate (ZCR) of an audio frame is the rate of sign-changes of the signal during the frame. The ZCR is defined according to the following equation:

$$Z(i) = \frac{1}{2}\sum_{i=2}^{n} |sgn[x_i] - sgn[x_{i-1}]| \quad (4)$$

where sgn(.) is denoted by:

$$sgn[x_i(n)] = \{\{1, x_i(n) \geq 0\}\{0, x_i(n) < 0\}\} \quad (5)$$

I. Mean Crossing Rate

MCR is a measure that reflects how many times the sign of two adjacent values in the signal crosses the mean. MCR is defined according the following equation:

$$MCR = \frac{\sum_{i=2}^{nclasses} |sgn[x_i - \mu] - sgn[x_{i-1} - \mu]|}{2} \quad (6)$$

where

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (7)$$

III. EXPERIMENTS AND RESULTS

A. Experiments

Experiments included comparative evaluations for a single statement in the dataset “Dogs are sitting by the door” using mel-cepstral and wavelet transformations both on continuous and discrete scales.

We used a 25ms speech window with mel–cepstral and 32ms window with wavelet features, due to specific decomposition structure. The mother wavelet chosen for continuous transform in signal decomposition was the Morlet wavelet. The mother wavelet chosen for discrete transform is Daubechies [9] with 28 filter length. The stride is set to 10ms for all extraction methods to ensure a fair comparison.

The Decision Tree Ensemble also termed the Random Forest method was used. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is the same as the original input sample size and the samples are drawn with replacement.

Quality of split is measured by “gini” denoting Gini Impurity. Max depth of the tree is set at 500. Maximum features to be used for prediction are nfeatures. The minimum number of samples required to be at a leaf node is set at 3. A split point at any depth will only be considered if it leaves at least 3 training samples in each of the left and right branches. The minimum number of samples required to split an internal node is set at 5. 400 estimators are used which denotes the number of trees in the forest. Model is kept similar to the three feature extraction methods to test the performance.

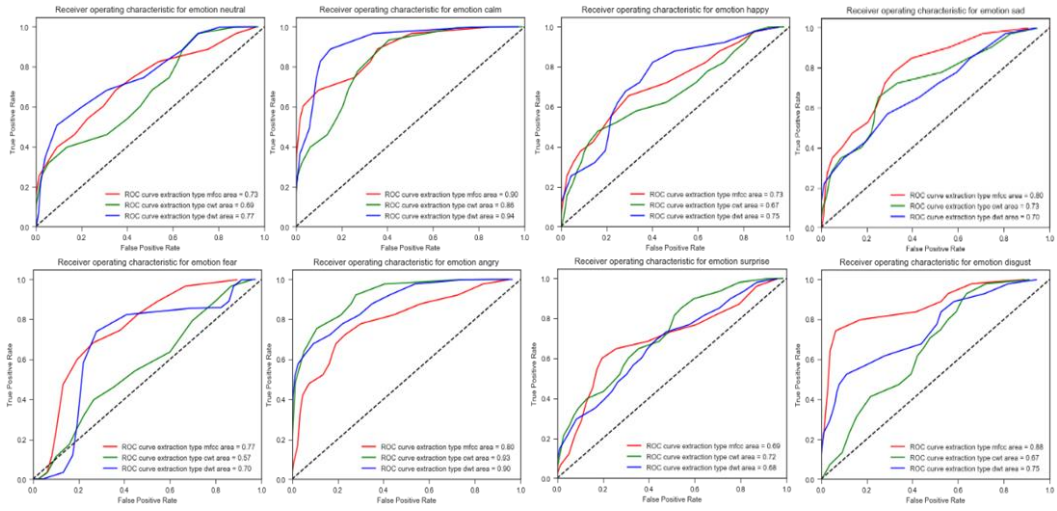


Fig. 1 Area under the curve of Receiver operating characteristics by the classifier for 8 emotions that are being studied

B. Performance

1) ROC curves:

The receiver operating characteristic (ROC) curve is a plot of True positive rates vs the False

positive rates. TPR corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

In other words, the higher the TPR, the fewer positive data points we will miss. FPR corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points. In other words, the higher the FPR, the more negative data points are misclassified. The performance of each emotion is measured by the area under the ROC-curve. To classify and compare each emotion we've used one vs all technique to get metric scores.(mention all scores are test scores)

Table 1: Comparative Table for the roc values by emotion

Emotion	**MFCC 128ms window width, 50-PC's**	**Continuous Wavelet Transform: 30-PC's**	**Discrete Wavelet Transform: 25-PC's**
neutral	**0.861868**	0.715714	0.833626
calm	0.916703	0.941538	**0.972857**
happy	0.709032	0.770968	**0.804839**
sad	0.736875	**0.828516**	0.731172
angry	0.858387	**0.943871**	0.895161
fear	**0.799231**	0.717033	0.735604
disgust	0.860775	**0.870641**	0.742608
surprise	0.694444	**0.723611**	0.687778

Table 2: Comparative Table for the accuracy values by emotion

Emotion	**MFCC 128ms window width, 50-PC's**	**Continuous Wavelet Transform: 30-PC's**	**Discrete Wavelet Transform: 25-PC's**
neutral	0.902778	0.902778	0.902778
calm	0.916667	0.888889	0.916667
happy	0.861111	0.861111	0.861111
sad	0.888889	0.888889	**0.902778**

angry	0.861111	**0.930556**	0.930556
fear	**0.902778**	0.888889	0.875
disgust	0.847222	0.833333	0.847222
surprise	0.833333	0.833333	**0.861111**

IV. DISCUSSION

From the Table 1 it is evident that wavelets outperform window fourier functions on more than half of inspected emotional classes. Although MFCC performed quite well considering their constant time-frequency behaviour.

We also tested the emotion recognition performance using a longer (>25ms) speech window in the MFCC calculation. They were found to be consistently worse for longer window durations.

During cross-validation analysis, it is found that in emotion “fear” MFCC consistently outperforms wavelet methods. It is now an open ended research where the reason behind the performance difference in one particular emotion can be studied.

In conclusion, despite the preliminary stage of our experimental setup in the field of non-linear speech emotion recognition, the results confirmed the hypothesis that wavelets can enhance the results of speech emotion recognition. Further work and improvements should incorporate the use of differential and acceleration (delta and delta-delta) coefficients.

REFERENCES

1. Mallat S. A wavelet tour of signal processing. San Diego: Academic Press 1999.
2. CBMS-NSF regional conference series in applied mathematics. in Daubechies Ingrid, Ten lectures on wavelets 1994.
3. O'shaughnessy D. Speech Communication: Human and Machine. NY: Addison-Wesley Publishing Company 1987.
4. Modic R. Comparative wavelet and MFCC speech recognition experiments on the Slovenian and English speechdat2. NOLISP. 2003.
5. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE. 2018;13(5):e0196391-e0196391.
6. Shegokar P, Sircar P. Continuous Wavelet Transform based Speech Emotion Recognition 2016.
7. Pattnaik , Sasweta M, Dash SK, Sabut . DWT-based feature extraction and classification for motor imaginary EEG signals. 2016 International Conference on Systems in Medicine and Biology (ICSMB). 2016.
8. Daubechies I. Orthonormal bases of compactly supported wavelets. Communications on Pure and Applied Mathematics. 1988;41(7):909-996.
9. Trang , Tran H, Nam , Huynh . 2015.